Computational approach to Word sense disambiguation in Tamil

Dr. S.Rajendran, Dr. M. Anand Kumar and Dr. Soman K.P
Amrita Vishwa Vidyapeetham,Coimbatore,India.
rajushush@gmail.com,m_anandkumar@cb.amrita.edu,kp_soman@amrita.edu

Native speakers of any natural language have an implicit knowledge of their language usually referred to as competence. Part of this tacit knowledge is concerned with interpretation, i.e. native speakers are capable of assigning an interpretation to any structure they generate. Native speakers not only assign an interpretation to every structure in their language, but also know that there are structures that may have more than one semantic interpretation. These structures are usually referred to as ambiguous structures. When an utterance has more than one interpretation, it is usually referred to as ambiguous. Ambiguity means that utterances may differ semantically but not phonetically, i.e. they differ in their interpretation but not in their form. Ambiguity may result from two homonyms occurring in the same structural position, as in the following example.

அவன் கால் பகுதியைச் சாப்பிட்டான் .
avan kaal pakutiyaic caappiTTaan
'He ate quarter of something'/He ate the leg part of something'

It may also occur when constituents in larger structures have more than one interpretation according to their internal structure and syntactic position.

வெள்ளை மருந்து குப்பி .
veLLai marundtu kuppi
'medicine bottle which is white in colour/a bottle with white medicine'

The fist one is called lexical ambiguity and the second structural ambiguity. Lexical ambiguity refers to the type of ambiguity those results from the occurrence of homonyms.

Here is in this paper we are concerned with the lexical ambiguity only.

1 Lexical ambiguity

The lexical ambiguity is very common. It includes, for example, the nouns such as *படி/paTi, குடி/kuTi, மடம்/maTam,* etc, verbs such as பிடி/*piTi, அறை/aRai, முடி/muTi, சுடு/cuTu*, etc and the adjectives such as *விரிந்த/virinta, குறைந்த/kuRainta, வெளுத்த/veLutta, கறுத்த/kaRutta,* etc. There are tests for establishing lexical ambiguity. One test is that for the word *கடினமான/kaTinamaana* there are two opposite words, *மெதுவான/metuvaana and எளிதான/eLitaana*. Consider the following example,

கடினமான மிட்டாயை கடிக்க முடியாது .
kaTinamaana miTTaayaik kaTikka mutiyaatu
'You cannot bite hard seet'

கடினமான சொல்லுக்கு பொருள் கூற இயலாது .
kaTinamana collukkup poruL kuuRa iyalaatu
'you caanot give meaning to hard word'

The reason for this ambiguity is that the word has more than one meaning. But it is not clear when there is only one word involved in ambiguity.

Though the noun படி/*paTi* and the verb படி/*paTi* have same spelling/pronunciation they are two different words. They are examples of homophones. One may wonder whether the noun கடி/*kaTi* and the verb கடி/*kaTi* are examples for homonyms or not. Doubt may arise whether the word mutal in *முதல் மாணவன்/mutal maaNavan* and *ஐந்து முதல்/aindtu mutal* are one and the same or not. To tell that one shows lexical ambiguity and the other is homonymous is not correct for all. This may be accidental.

There are three basic types in lexical ambiguity: category ambiguity, ambiguity due to homography and ambiguity due to ploysemy.

1.1 Category ambiguity

Category ambiguity is the most straight forward type of lexical ambiguity. This happens when a given word may be assigned to more than one grammatical or syntactic category as per

context. One can find a number of such examples in Tamil. For example the word paccai ‘green’ can be both noun as well as adjective. Similarly the word *cuTu* can be both verb as well as an adverb. *kaTi* could be both verb as well as noun. As shown in the following example *paTi* can be noun as well as a verb denoting two different meanings and thus shows ambiguity.

அவன் தந்தை அவனிடம் நீ நன்றாக படி என்று கூறினார் .
avan tantai avaniTam nii nanRaakap paTi eRu kuuRinaar
‘His father asked to study well’

அவன் மாடி படி வழியாக மேலே ஏறினான் .
Avan maaTip paTi vaziyaaka meelee eeRinaan.
‘He climbed up through stair case’

The words like *meelee* and *kiizee* could be adverbs and postpositions.

அவன் மேலே இருக்கிறான் .
avan meelee irukkiRaan. (noun)
He is at the top

அவன் மேலே சென்றான் .
avan meelee cenRaan (adverb)
He went up

அவன் மேஜை மேலே நிற்கின்றான் .
Avan meecai meelee niRkiRaan. (postposition)

Category ambiguities can be often be resolved by morphological inflection. For example aTi in *avan aTikkiRaan* ‘he is beating’ is a verb and *அடி/aTi* in *அவனால் அந்த அடியைத் தாங்க முடியவில்லை/avanaal andta aTiyait taangka muTiyavillai* ‘He could not bear that beating’ is noun. Frequently ambiguity can be resolved by syntactic parsing. However, the problem increases when several categorical ambiguous words occur in the same sentence, each requiring being resolved syntactically.

1.1.1 Categorical ambiguity due to historical functional reorganization

The inflected forms of nouns or verbs will denote different word category or functional category due to historical meaning change. For example many of the postpositions in Tamil are historically the inflected forms of verbs. The inflected forms இருந்து/*irundtu* 'from' *, பற்றி/paRRi* 'about'*, குறித்து/kuRittu* 'about'*, ஒட்டி/oTTi* 'about', கொண்டு/*koNTu* 'by (means of)'*, வைத்து/vaittu* 'by (means of)'*, சுற்றி/cuRRi* 'around'*, நோக்கி/ndookki* 'towards'*, முந்தி/mundti* 'before'*, விட/viTa* 'than' *,* and கூட/*kuuTa* 'along with' are the inflected forms of the verb *iru* 'be', பற்று/*paRRu* 'catch', குறி/*kuRi* 'aim', ஒட்டு/*oTTu* 'stick', கொள்/*koL* 'have', வை/*vai* 'keep', சுற்று/*cuRRu* 'go aroung', நோக்கு/*ndookku* 'look at', *mundtu* 'over take', *viTu* 'leave', and *kuuTu* 'assemble' respectively.

1a. அவன் வீட்டில் **இருந்து** வெளியேறினான். (முன்னுருபு)

avan viiTTilirundtu veLiyeeRinaan

'He went out from the house'

1b. அவன் வீட்டில் **இருந்து** வந்தான் (வினைத்திரிபு வடிவம்)

avan viiTTil irundtu vandtaan

'He was in the house (habitually/condinuously)'

2a. அவன் அவளைப் **பற்றி** பேசினான். (முன்னுருபு)

avan avaLaip paRRi peecinaan

'He talked about her'

2b. அவன் அவள் கையைப் **பற்றி** முத்தமிட்டான். (வினைத்திரிபு வடிவம்)

avan avaL kaiayip paRRi muttamiTTaan

'He catch hold of her hand and kissed it'

3a. அவன் அவளைக் **குறித்து** பேசினான். (முன்னுருபு)

avan avaLaik kuRittu peecinaan

‘He talked about her’

3b. அவன் அவள் சொல்வதைக் **குறித்து** வந்தான். (வினைத்திரிபு வடிவம்)

avan avaL colvataik kuRittu vandtaan

‘He was noting down what she was telling’

4a. அவன் அந்த தலைப்பை **ஒட்டி** பேசினான். (முன்னுருபு)

avan andta taippai oTTi peecinaan

‘He talked about that title’

4b. அவன் போஸ்டர் **ஒட்டி** பிழைக்கின்றான். (வினைத்திரிபு வடிவம்)

avan poosTar oTTi pizaikkinRaan

‘He eke his livelihood by pasting posters’

5a. அவன் கத்தி **கொண்டு** அதை வெட்டினான். (முன்னுருபு)

avan katti koNTu atai veTTinaan

‘He cut it with a knife’

5b. அவன் பென்சிலைச் சீவிக் **கொண்டு** பேசினான். (வினைத்திரிபு வடிவம்)

avan pencilaic ciivik koNTu peecinaan

‘He was speaking while sharpening the pencil’

6a. அவன் கத்தி **வைத்துப்** பழம் வெட்டினான். (முன்னுருபு)

avan katti vaittup pazam veTTinaan

‘He cut the fruit with the knife’

6b. அவன் பணம் **வைத்துக்** கொண்டு சூதாடினான். (வினைத்திரிபு வடிவம்)

avan paNam vaittuk koNTu cuutaaTinaan

‘He gambled by keeping the money at hand’

7a. அவன் வீட்டைச் **சுற்றி** மரங்கள் நிற்கின்றன. (முன்னுருபு)

avan viiTTaic cuRRi marangkaL ndiRkinRana

‘The house is surrounded by the tree’

7b. அவன் அவளையே **சுற்றி** வருகின்றான். (வினைத்திரிபு வடிவம்)

avan avaLaiyee cuRRi varukinRaan

‘He is going after her’

8a. அவன் அவளை **நோக்கி** நடந்தான். (முன்னுருபு)

avan avaLai ndookki ndaTandtaan

He went towards her’

8b. அவன் அவள் முகத்தை **நோக்கிச்** சிரித்தான். (வினைத்திரிபு வடிவம்)

avaL mukattai ndookkic cirittaan

‘He smiled looking after her face’

9a. அவன் அவளுக்கு **முந்தி** அங்கு வந்தான். (முன்னுருபு)

avan avaLukku mundti angku vandtaan

‘He came there before her’

9b. அவன் அவளை **முந்தி** நடந்துகொண்டிருந்தான். (வினைத்திரிபு வடிவம்)

avan avaLai mundti ndaTandtukoNTirundtaan

‘He is walking overtaking her’

10a. அவன் அவளை **விட** நல்லவன். (முன்னுருபு)

avan avaLai viTa ndallavan

‘He is better than her’

10b. அவன் அவளை **விட** விருப்பவில்லை. (வினைத்திரிபு வடிவம்)

avan avaLai viTa virumpavillai

‘He does want to leave her’

11a. அவன் அவள் **கூட** வந்தான். (முன்னுருபு)

avan avaL kuuTa vandtaan

‘He came with her’

11b. அவன் அவர்களுடன் **கூட** விரும்பினான்

avan avarkaLuTan kuuTa virumpinaan

‘He wanted to gather together with them’

The word *என்று/enRu* which is the inflected from the verb en ‘say’, show two different grammatical functions thus showing ambiguity.

a.அவன் நல்லவன் என்று நினைத்தேன் (என்று/enRu functions as complementizer)
avan ndallavan enRu ndinaitteen
b. அவன் திடீர் என்று வந்தான் (என்று/enRu functions as adverbializer)
avan tiTiir enRu vantaan

The word என்று/enRu is having homographic relation with the inflected verbal form என்று/enRu.

avan enRu varukiRaan
அவன் என்று வருகிறான்
‘When does he come?”

The inflected verbal forms which can be analyzed verb root+um (future suffix) can be interpreted at least in two ways.

Atu naaLai varum

‘It will come tomorrow’

Atu varum naaL enakkut teriyaatu

‘I don’t know the date of its coming’

The inflected verbal form which can be analyzed as verb + tense + அது/atu can be interpreted in three ways.

atu vandtatu ‘it came’ (*vandtatu* is the finite verbal form)

atu vandtatu enakkut teriyaatu (*vandtatu* is the gerundival form)

“I did know that it had come’

andta ceytittaaL neeRRu vandtatu (*vantatu* is the participial noun)

‘That newspaper is yesterday’s one’

The ambiguity can be resolved by selection restriction, context, collocation, co-occurrence, etc.

1.2 Ambiguity due to Homography

If two entirely different words having same form have different meanings, the ambiguity arises due to homography. In the following examples the word அட்டை/*aTTai* shows homography; அட்டை/*aTTai* can denote ‘leach’ as well as ‘binding’.

அவன் புத்தகத்தின் அட்டையைக் கிழித்து எறிந்தான் .
Avan puttakattin aTTaiyaik kizittu eRintaan
‘He tore away the binding’
அவன் அட்டையைக் கொன்றான் .
Avan aTTaiyaik konRaan
‘He killed the leach’

Sometimes among the homographs, the use of one may be greater than the other. In that case the ambiguity can be resolved on the basis of text. This is done by setting aside the unusual meaning form the dictionary unless it is required for translation.

1.2.1 Homography in inflected words

Homogrphy may arise when two different words get inflected differently resulting in homographic forms. The homography can be resolved by different morphological analysis. The following examples will reveal this.

அவன் கடலை தின்று மகிழ்ந்தான் .
avan kaTalai tinRu makizntaan
'He enjoyed eating pea nut'

அவன் கடலை கண்டு மகிழ்ந்தான் .
avan kaTalai kaNTu makizntaan
'He enjoyed seeing the sea'

In the first sentence the noun கடலை/*kaTalai* denotes 'pea nut' and in the second case கடலை/*kaTalai* has to be analysed as கடல்/*kaTal* + ஐ/*ai* (accusative case marker) and interpreted as *kaTal* 'sea'.

As in the case of the following example, the inflected word of one type of morphological analysis resembles an inflected word form of another morphological analysis, there by showing homogrphy.

அவன் (துணி) நெய்தான் .
avan (tuNi) ndeytaan.
'He weaved (cloth)'

அவன் நெய்தான் விரும்புகிறான் .
avan ndeytaan virumpukiRaan
'He likes gee only'

In the first sentence the word நெய்தான்/*ndeytaan* has to be interpreted after analyzing it into நெய் + த்/*ney*+*t* (past tense)+ஆன்/*aan* (third person masculine singular) and in the second sentence *neytaan* has to be interpreted as நெய்/ndey 'ghee' + தான்/taan 'only'. Even the two root words *ney* 'weave' and *ney* 'ghee' are homogrpahs showing categorical lexical ambiguity.

1.3 Ambiguity due to ploysemy

If a word has two or more meanings it can be said that the ambiguity is due to polysemy. Polysemy expresses extension of meaning. The polysemous words may express new meaning by metaphoric and metonymic extensions. For example the word கிளை/*kiLai* 'branch' may denote branch of a tree as well as a branch of a bank. நட/*ndaTa* can denote the action of walking as well as happening or functioning of something.

அவன் தினமும் காலையில் பள்ளிக்கு நடந்து செல்கிறான் .
avan tinamum kaalaiyil paLLikku ndaTantu celkinRaan
He goes to school daily by walking'

அந்த நிறுவனம் நன்றாக நடந்துக்கொண்டிருக்கின்றது
Anta niRuvanam nanRaaka ndaTandtukoNTirukkinRatu
'That organization is functioning well'

அந்த தியேட்டரில் சினிமா நடந்துகொண்டிருக்கின்றது .
Anta tiyeeTTaril cinimaa ndaTandtukoNTirukkinRatu
'A cinema is running in the theatre'

ஓடு/*ooTu* can denote human action of running as well flow of a river.

அவன் விரைவாக ஓடுகிறான் .
Avan viraivaaka ooTukiRaan
'He is running fast'

தஞ்சாவூர் வழியாகக் காவேரி ஓடுகிறது .
Tanjaavuur vaziyaakak kaaviriyaaRu ooTukiRatu
'The river Kaviri runs through Thanjavur'

கண்/*kaN* may denote the eye of animate beings as well as the eye like spot in the coconut.

அவன் தன் கண்களை மூடினான் .
avan tan kaNkaLai muuTinaan

‘He closed his eyes’

தேங்காய்க்கு மூன்று கண்கள் உண்டு .
Teengkaaykku muunRu kaNkaL uNTu
‘there are three eye like spots in the coconut’

Similarly many words denoting body parts show polysemous extension thereby denoting ambiguity. For example வாய்/*vaay* can be a human mouth as well as the mouth of a bottle; கால்/*kaal* can be a human leg or leg of furniture. In the following example the ambiguity is due to metonymic extension.

ஊர் சிரித்தது
uur cirittatu
‘The people (of the village) laughed’

Here in this sentence ஊர் /uur ‘people’ is used as a metonymic extension of ஊர் /uur ‘village’.

In the following sentence the word கேள்/*keeL* denotes both the perception through ears as well as ‘asking’.

ராதை ராஜா கேட்டதை அவனிடம் கூறினாள் .
raatai raajaa keeTTatai avaniTam kuuRinaaL
‘Radha told him what Raja has asked her’
‘Radha told him that Raja has heard that’

This sentence is ambiguous giving at least two interpretations.

ராதை ராஜா தான் காதால் கேட்டதை அவனிடம் கூறினாள் .
Raatai raajaa tan kaataal keeTTatai avaniTam kuuRinaaL
‘Radha told him what Raja has heard’

ராதை ராஜா வினவியதை அவனிடம் கூறினாள் .
Raatai raajaa vinaviyatai avaniTam kuuRinaaL
‘Radha told him that Raja had asked her’

As for machine translation is concerned both the homography and ploysemy will be treated alike as the aim is to finding out the meaning by context. The homographs belonging to different grammatical categories can be resolved as explained before. But they belong to same grammatical categories syntactic parsing may not be enough. One common approach is to assign semantic features such as 'human', 'female', 'liquid' etc and to specify which features are compatible in the given syntactic constructions, via selection restrictions. For example it might be specified that the verb குடி/*kuTi* 'drink' have an 'animate' subject.

There are difficulties in finding semantic features that can be used consistently and specifying the selection restriction for nouns and verbs based on these features. Even then these are widely used in Machine translation system often in combination with case roles. But the semantic features cannot solve all the problems, even in situations for which they have been devised. For example let us take the word அட்டை/*aTTai*. As we have found out it is used in the senses of 'binding' and 'leach'. These two meanings can be differentiated explaining the relevant co-occurrence restrictions we find out in the following sentences in which அட்டை/*aTTai* is used.

புத்தகத்தின் அட்டை கிழிந்துவிட்டது .
Puttakattin aTTai kizintuviTTatu
'The binding of the book is torn'

அட்டை ஊர்ந்து செல்கிறது .
aTTai uurndtu celkinRatu
'The leach is crawling'

The verbs like கிழி/*kizi* will take the objects like அட்டை/*aTTai* which can be torn as their subjects and the verbs like ஊர்ந்துசெல்/uurndtucel takes objects like அட்டை/*aTTai* which can crawl.

2. Word Sense Disambiguation

Word sense disambiguation is finding out the correct senses of a word in its context though computational methodology. Word sense disambiguation (WSD) is considered as AI-complete problem. This is equivalent to very hard problem in AI.

2.1 External knowledge resources.

The important component of word sense disambiguation is knowledge. Knowledge resources gives data which is needed to relate words with meanings. They include unlabeled or annotated corpora of texts, machine readable dictionaries (MRDs), thesauri, glossaries, ontologies and others. Structured resources includethesaurus, Machine readable Dictionaries (MRDs), and ontologies. Unstructured resources include corpora, raw corpora, sense-annotated corpus, collocational resources and other resources.

2.2 Representation of context

As text is an unstructured source of information, to make it a suitable input to an automatic method it is usually transformed into a structured format. To this end, a preprocessing of the input text is usually performed, which typically (but not necessarily) includes the following steps:

tokenization, a normalization step, which splits up the text into a set of tokens (usually words);

part-of-speech tagging, consisting in the assignment of a grammatical category to each word (e.g. "அந்த/anta_DT viiTu/வீடு_NN azakaaka/அழகாக_ADV irukkiRatu/இருக்கிறது_VBD," where DT, NN, ADV, and VBD are tags for determiners, nouns, adverbs and verbs respectively);

lemmatization, that is, the reduction of morphological variants to their base form (மரங்கள்/marangkaL > மரம்/maram, வந்தான்/vandtaan > வா/vaa);

chunking, which consists of dividing a text in syntactically correlated parts (e.g., [அந்த வீடு/anta viiTu]NP [அழகாக இருக்கிறது/azakaaka irukkiRatu]VP, respectively the noun phrase and the verb phrase of the example); and

parsing, whose aim is to identify the syntactic structure of a sentence (usually involving the generation of a parse tree of the sentence structure).

2.3 Choice of a Classification Method

The final step is the choice of a classification method. Most of the approaches to the resolution of word ambiguity stem from the field of machine learning, ranging from methods with strong supervision, to syntactic and structural pattern recognition approaches. We can broadly distinguish two main approaches to WSD:

supervised WSD: these approaches use machine-learning techniques to learn a classifier from labeled training sets, that is, sets of examples encoded in terms of a number of features together with their appropriate sense label (or class);

unsupervised WSD: these methods are based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context.

We further distinguish between knowledge-based (or knowledge-rich, or dictionary based) and corpus-based (or knowledge-poor) approaches. The former rely on the use of external lexical resources, such as machine-readable dictionaries, thesauri, ontologies, etc., whereas the latter do not make use of any of these resources for disambiguation.

Finally, we can categorize WSD approaches as token-based and type-based. Token based approaches associate a specific meaning with each occurrence of a word depending on the context in which it appears. In contrast, type-based disambiguation is based on the assumption that a word is consensually referred with the same sense within a single text. Consequently, these methods tend to infer a sense (called the predominant sense) for a word from the analysis of the entire text and possibly assign it to each occurrence within the text. Notice that token-based approaches can always be adapted to perform in a type-based fashion.

The methodology we have chosen works in two phases. At the beginning of the training phase the system identifies the ambiguous words in the text using the ambiguous words list. For each occurrence of an ambiguous word a context is identified from the text. Different researchers have used different sizes of contexts, i.e. different window sizes. The window size influences the performance of the system considerably. A larger window size brings in more unrelated words to the context, while a smaller window size misses some important collocations. The frequent words also called *stop-words* are removed from the window. The context is then morphologically analysed to get the root form of the contextual words, which are used, in clustering. Each context is represented as a context or occurrence vector. Using the K-means approach these vectors are collected into different clusters. Collocations are then collected from these clusters automatically and senses are assigned to these collocations by human-annotators thereby developing a sense-collocation dictionary.

2.4 Significance of Collocations

Collocations are nearby words that strongly suggest the sense of the ambiguous word in a given sentence. In general, the term collocation refers to a quantifiable position specific relationship between two lexical items. Collocations encode information about words that are semantically *very* close to the ambiguous word. For example consider the word ‘plant’ having the two following senses.

plant1 – a living organism

plant2 – a manufacturing place

Collocations for the word ‘plant’ will include the words that frequently occur with ‘plant’ in texts. Some typical collocations for the above two senses of ‘plant’ are listed below.

plant1 – growth, height, flower, fruit, species, leaves 2.3

plant2 – car, union, equipment, assembly, nuclear, job, worker 2.4

Generally collocations are sensitive to the distance from the ambiguous word. Words that are at a short distance from the ambiguous word strongly indicate its sense; as the distance increases the relevance of the word to that of the ambiguous word decreases.

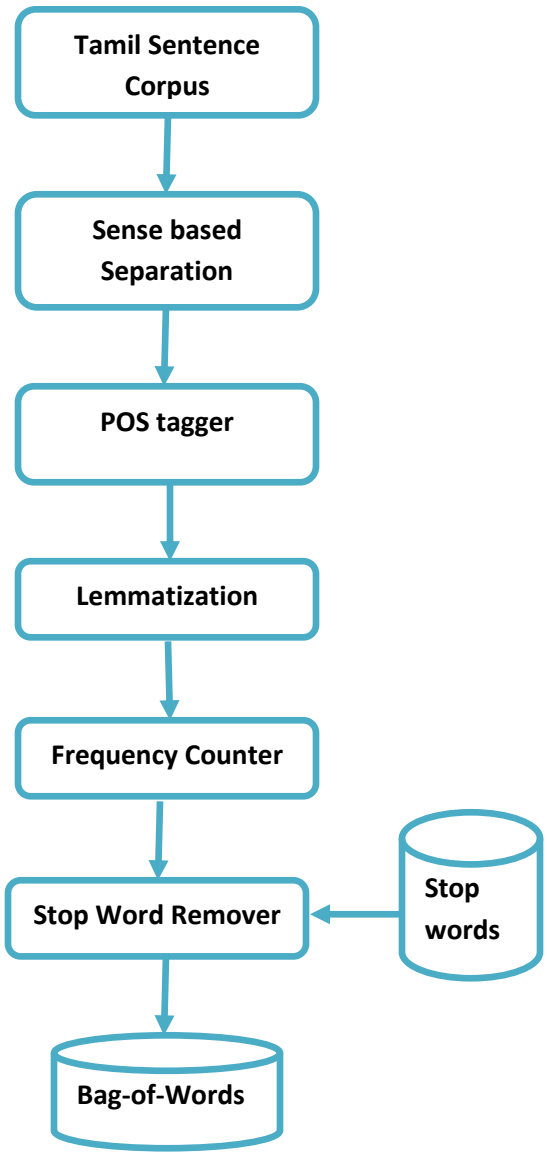
The distance is commonly referred to as window size. However, there is no fixed distance from the ambiguous word within which a word can be considered as a collocation. Sometimes, collocations may even occur far away from the ambiguous word.

2.5. Building of the present system.

The system has two phases: the training phase and the testing phase. In the training phase the bags of words or cluster of words occurring along with the ambiguous word will be obtained. The number of bags or cluster depends on the number of ways a particular word is ambiguous. For example the ambiguous word *aTTai* has two senses, and then we have to collect two bags of words or cluster of words.

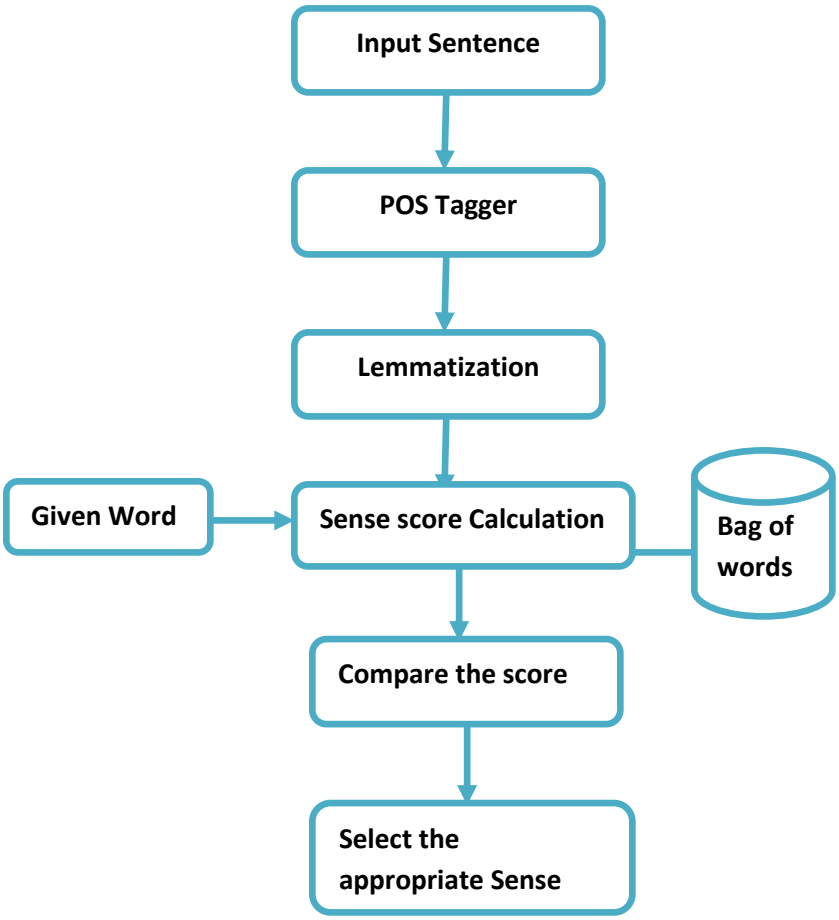
The ambiguous words will be identified and the sentences in which they occur will be extracted from the corpus. For example, the Tamil word அட்டை/*aTTai* at least has two senses: அட்டை/*aTTai* 'leach' and அட்டை/*aTTai* 'binding of a book. From the corpus the sentences with the target ambiguous word (அட்டை/*aTTai*) will be extracted and manually separated into two sets. Each set contains twenty sentences each. The sentences will be post-tagged and lemmatized. The occurrence of words along with the ambiguous word will be counted for each set of sentences. The functional words or words which do not contribute to the meaning of the ambiguous word are known as stop words. Such words need to be removed from sets of contextual words. After the removal of stop-words we will get a finite number of bags or cluster of words each belonging to the different senses of the ambiguous word. Now the system is ready for testing phase

Diagram representing the training phase



In the testing phase, the new sentence with the ambiguous word will be sense marked using the learning phase, i.e. using the bag of words. The input sentences will be post-tagged and lemmatized. The ambiguous word will be calculated sense score using the bag of words identified for the different senses. Using the sense score calculation, the appropriate sense of the ambiguous word will be selected.

Diagram representing the testing phase



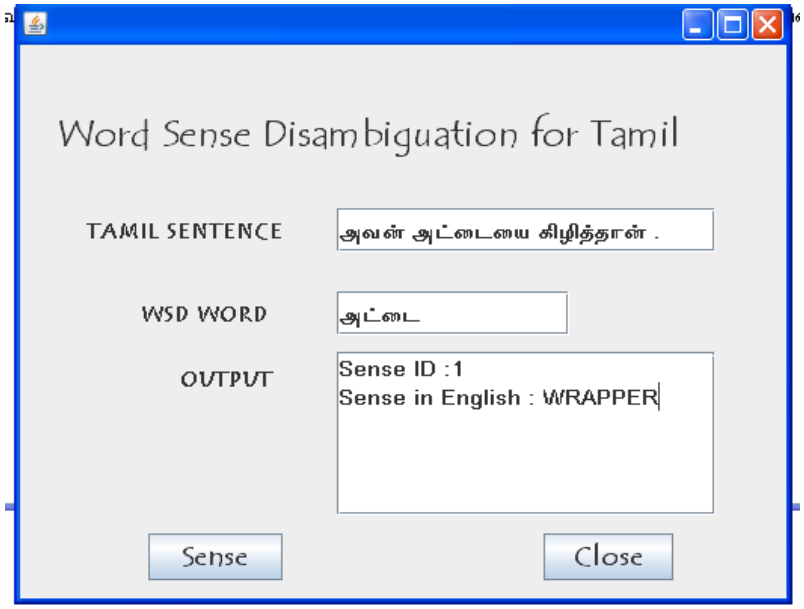
Using this method the ambiguity of the following ambiguous words such as *மாலை maalai, நூல்/nduul, விலங்கு/vilangku* can be disambiguated. The following table shows the bag of words associated with each senses of the ambiguous words.

Word	Meaning	Word occurring together
மாலை1	evening	காலை, பகல், சூரியன், மறை, நேரம், நாள், மாதம், மதியம், இரவு
மாலை2	garland	பூ, மணம், ரோஜா, மல்லிகை, மலர், பூஜை, மகளிர், கொன்றை
நூல்1	thread	இழை, தறி, சாயம், நெய், கதர், நெசவு, ஊடு, பாவு,

		துணி, ஊசி
நூல்2	book	ஆசிரியர், பதிப்பகம், அச்சு, பக்கம், தாள்
விலங்கு1	animal	காடு, சிங்கம், புலி, தோல், யானை, தந்தம், நாய், பூனை
விலங்கு2	Hand-cuff	குற்றவாளி, கோர்ட், போலீஸ், கைதி, விசாரணை, தண்டனை

Graphical User interface

The following is the graphical user interface for the system. The ambiguous words can be disambiguated using the GUI.



3 Conclusion

There are various methods for WSD. It can be said that for WSD unsupervised approach is economical than supervised method. The knowledge based approaches are welcome thing. The clustering method is a non-supervised method. The present approach is giving encouraging result.

References

இராசேந்திரன், ச. 1999. பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும் [Classification into semantic domains & Thesaurus]. புலமை தொகுதி 25, எண்2, டிசம்பர், 1999, 47-66.

இராசேந்திரன், ச. 2001 தற்காலத் தமிழ்சொற்களஞ்சியம். தமிப்பல்கலைக் கழகம், தஞ்சாவூர்.

இராசேந்திரன், ச. 2006. தமிழ் மின்சொற்களஞ்சியம். தமிழ்ப்பல்கலைக் கழகம், தஞ்சாவூர்.

ARUN, M.C. “Evaluation of Word Sense Disambiguation methods applied on Tamil Corpus. PhD thesis.

BASKARAN, S. 2002. Semantic analyser for word sense disambiguation. MS thesis. Madras Institute of Technology, Anna University, Chennai 2002.

BASKARAN SANKARAN, VAIDEHI V, 2002, "Role of Collocations and Case-Markers inWord Sense Disambiguation: A Clustering-Based Approach". In Proceedings of IEEE International Symposium on Natural Language Processing and Knowledge Engineering 2002. Vol. I. pp. 625-630. Hammamet, Tunisia.

BASKARAN S, VAIDEHI V, 2003, "Collocation based Word Sense Disambiguation using Clustering for Tamil". Communicated to International Journal of Dravidian Linguistics, Thiruvananthapuram, India.

IDE, N. AND V´ERONIS, J. 1998. Word sense disambiguation: The state of the art. Computat. Ling. 24, 1, 1–40.

NAVIGLI, R. 2009. Word Sense Disambiguation: A Survey. ACM Computing Surveys, Vol 41, No.2, Article 10, Publication date February 2009.

PUSTEJOVSKY, J. 1995. The Generative Lexicon. MIT Press, Cambridge, MA.

SOWA, J.F. 1984. Conceptual Structures: Information Processing in Mind and Machine. Reading, Massachusetts: Addison-Wesley Publishing Company.

YAROWSKY, D. 1992. Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. In Proceedings of the 14th International Conference on Computational Linguistics (COLING, Nantes, France). 454–460.

YAROWSKY, D. 1993. One sense per collocation. In Proceedings of the ARPA Workshop on Human Language Technology (Princeton, NJ). 266–271.

YAROWSKY, D. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM). 88–95.

YAROWSKY, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Cambridge, MA). 189–196.

YAROWSKY, D. 2000. Hierarchical decision lists for word sense disambiguation. Comput. Human. 34, 1-2, 179–186.

YAROWSKY, D. AND FLORIAN, R. 2002. Evaluating sense disambiguation across diverse parameter spaces. J. Nat. Lang. Eng. 9, 4, 293–310.